ECE 598MR: Statistical Learning Theory

Maxim Raginsky

Homework 2

Assigned October 2, 2014; due October 14, 2014

Note: natural logarithms are used throughout, unless stated otherwise.

1. **VC classes.** Prove the following statements.

   (a) Let $\mathscr{C}$ and $\mathscr{C}'$ be two classes of subsets of some feature space X. Suppose that $\mathscr{C} \subseteq \mathscr{C}'$, meaning that if $C \in \mathscr{C}$, then $C \in \mathscr{C}'$ as well. Prove that $V(\mathscr{C}) \le V(\mathscr{C}')$.

   (b) Let $\mathscr{C}$ be a *finite* class of subsets of X. Prove that $V(\mathscr{C}) \le \log_2 |\mathscr{C}|$.

   (c) Let X be a *finite* feature space. For a given $k \le |\mathsf{X}|$, consider the class $\mathscr{F}_k$ of binary-valued functions $f : \mathsf{X} \to \{0,1\}$, such that $\big|\{x \in \mathsf{X} : f(x) = 1\}\big| = k$. Find $V(\mathscr{F}_k)$.

2. **An alternative to ERM.** Searching for an empirical risk minimizer in an infinite function class $\mathscr{F}$ may not always be feasible. Let's consider the following alternative procedure that reduces to searching over a *finite* subclass of $\mathscr{F}$. We will be looking at a binary classification problem, so let $\mathscr{F}$ be a class of functions $f : \mathsf{X} \to \{0,1\}$, where X is some feature space. Given a training sample $\{(X_i, Y_i)\}_{i=1}^n$ from an unknown probability distribution P on $\mathsf{X} \times \{0,1\}$, we carry out the following two-step procedure:

   - Pick some $m < n$. Let $\mathscr{B}_m$ be the set of all binary strings in $\{0,1\}^m$ of the form

     $$b^m(f) = (b_1(f), \dots, b_m(f)) = \big(f(X_1), \dots, f(X_m)\big) \tag{1}$$

     for some $f \in \mathscr{F}$. For each $b \in \mathscr{B}_m$, pick one $f \in \mathscr{F}$ such that (1) holds. Let $\widehat{\mathscr{F}}_m$ denote the (finite) set of all such f's. Note that this is a random set, since it depends on the sample $(X_1, \dots, X_m)$.

   - Compute

     $$\widehat{f}_n = \operatorname*{arg\,min}_{f \in \widehat{\mathscr{F}}_m} \frac{1}{n-m} \sum_{i=m+1}^{n} \mathbf{1}_{\{f(X_i) \neq Y_i\}}.$$

     This will be our actual classifier.

   What we have done is split the original training sample into two subsamples, used the first subsample to extract a finite subclass of $\mathscr{F}$, and then performed ERM over this subclass on the second subsample. We will now analyze the classification error of $\widehat{f}_n$.

(a) Let

$$\tilde{f}_m = \operatorname*{arg\,min}_{f \in \widehat{\mathscr{F}}_m} L(f)$$

be the best classifier in $\widehat{\mathscr{F}}_m$. Note that this is a random object, since it depends on the random set $\widehat{\mathscr{F}}_m$. Prove that

$$L(\widehat{f}_n) - L(\tilde{f}_m) \le 8\sqrt{\frac{\log|\mathbb{S}_m(\mathscr{F})|}{n-m}} + \sqrt{\frac{2\log(2/\delta)}{n-m}} \tag{2}$$

with probability at least $1-\delta/2$, where $\mathbb{S}_m(\mathscr{F})$ is the mth shatter coefficient of $\mathscr{F}$.
Hint: Use the fact that $\widehat{f}_n$ is a solution of an ERM problem over the second subsample, add and subtract appropriate empirical quantities, and then apply the Finite Class Lemma.

(b) Observe that

$$\begin{aligned} L(\tilde{f}_m) - L^*(\mathscr{F}) &\le \sup_{f,f' \in \mathscr{F}:\, b^m(f)=b^m(f')} \left|L(f) - L(f')\right| \\ &\le \sup_{f,f' \in \mathscr{F}:\, b^m(f)=b^m(f')} \mathbb{P}\left[f(X) \ne f(X')\right] \\ &\le \sup_{A \in \mathscr{A}} \left| P(X \in A) - \frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\{X_i \in A\}} \right|, \end{aligned}$$

where $\mathscr{A}$ is the class of all sets of the form $\{x \in \mathsf{X} : f(x) \ne f'(x)\}$ for all pairs $f, f' \in \mathscr{F}$. Use this to prove that

$$L(\tilde{f}_m) - L^*(\mathscr{F}) \le C\sqrt{\frac{V(\mathscr{A})}{m}} + \sqrt{\frac{2\log(2/\delta)}{m}}$$

with probability at least $1-\delta/2$, where $C > 0$ is an absolute constant. (It is not hard to show that $V(\mathscr{A}) \le 4V(\mathscr{F})$ — you don't have to do this.)

(c) Finally, use parts (a)–(b) to prove that

$$L(\widehat{f}_n) - L^*(\mathscr{F}) \le 8\sqrt{\frac{\log|\mathbb{S}_m(\mathscr{F})|}{n-m}} + C\sqrt{\frac{V(\mathscr{A})}{m}} + \sqrt{\frac{2\log(2/\delta)}{n-m}} + \sqrt{\frac{2\log(2/\delta)}{m}}$$

with probability at least $1-\delta$.

3. **Surrogate loss bound for a sigmoidal classifier.** Let the feature space X be a subset of $\mathbb{R}^d$. Consider the class $\mathscr{F}_R$ of functions f the form

$$f_w(x) \triangleq \tanh\left(\langle w, x\rangle\right), \tag{3}$$

where w runs over all vectors in $\mathbb{R}^d$ with $\|w\| \le R$. Each such f induces a classifier $g_f(x) = \operatorname{sgn} f(x)$. A classifier of this kind first computes a weighted sum of the features, then passes it through a smooth nonlinear function, and then computes the sign of the resulting value. The hyperbolic tangent is an example of a *sigmoidal function* (where "sigmoidal" is a fancy term for "S-shaped" — look at the graph of $u \mapsto \tanh u$). The transformation in (3) is a simple model of a nonlinear neuron.

Let φ be a surrogate loss function satisfying the assumptions of Theorem 3 in the lecture notes on Binary Classification, Part 1. Let $\widehat{f}_n \in \mathscr{F}_L$ be a function generated by an arbitrary learning algorithm on the basis of an i.i.d. sample $\{(X_i, Y_i)\}_{i=1}^n$ from an unknown probability distribution P on $\mathsf{X} \times \{-1, +1\}$. Prove that

$$L(\widehat{f}_n) \le A_{\varphi,n}(\widehat{f}_n) + 8RM_\varphi \sqrt{\frac{\mathbb{E}[\|X\|^2]}{n}} + B\sqrt{\frac{\log(1/\delta)}{2n}}$$

with probability at least $1 - \delta$, where M_φ and B are defined in Theorem 3. Recall that $L(f)$ is our shorthand for the error probability $\mathbb{P}[\operatorname{sgn} f(X) \neq Y]$.